Dirección de Desarrollo Digital

Unidad De Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

PREDICCIÓN DELITO BUCARAMANGA

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Desarrollo Digital - Unidad de Científicos de Datos • Dirección de Seguridad Justicia y Gobierno
Sector	Justicia
Tecnologías utilizadas	Python
Fuentes de datos	Delitos enviados por la Alcaldía de Bucaramanga

Contenido

1. Presentación

En este proyecto se desarrollaron modelos de modelo de aprendizaje de máquina para predecir delitos en Bucaramanga de manera semanal y por división geográfica. Este informe presenta el proceso que incluye estadísticas descriptivas de datos, predicciones de aprendizaje de máquina y un tablero de visualización de delitos históricos y predicciones.

This project developed machine learning models to predict crime in the city of Bucaramanga by week and geographical sectors. This report presents descriptive statistics, machine learning predictions and the dashboard to show prediction results and historical crime data.

2. Objetivos del proyecto

2.1. General

Desarrollar un modelo de predicción del delito que permita mejorar, si posible, los resultados obtenidos por el modelo realizado en 2020 en Bucaramanga y mostrar los resultados en una herramienta de visualización.

2.2. Específicos

1. Desarrollar un nuevo modelo de predicción del delito en Bucaramanga
2. Adecuar y utilizar las bases de datos enviadas por la Alcaldía de Bucaramanga y Dirección de Justicia, Seguridad y Gobierno en el modelo de predicción
3. Actualizar y, si necesario, modificar la herramienta de visualización de resultados elaborada en 2020

3. Metodología

La metodología se divide en las estadísticas descriptivas, las predicciones de aprendizaje de máquina y en el tablero de visualización.

3.1. Estadísticas descriptivas

- ***Índice de Moran***

El índice de Moran es una prueba que mide la existencia de autocorrelación espacial entre regiones contiguas. Es decir, mide si puede existir un efecto de cambios de una variable en una región frente a la medición de la misma variable en regiones vecinas. El Índice de Morán no es más que el coeficiente resultante de una regresión lineal de la variable de análisis en una región con respecto a esa misma variable en regiones contiguas. Las regiones contiguas pueden tener diferentes niveles de rezagos. Si una región B comparte frontera con otra (A), tiene un rezago de nivel 1 con A. Si una región C no comparte directamente con A pero sí con B, es de nivel dos con respecto a A. Así sucesivamente puede haber rezagos de más niveles.

La Ecuación 1 muestra la forma de cálculo del índice de morán donde y es la variable, $\bar{y}$ es el valor promedio de la variable, i y j son índices de dos regiones ($i \neq j$) y w corresponde a una matriz cuadrada de dimensiones nxn donde n es la cantidad total de regiones. Esta mide la relación espacial existente entre las regiones. Esta matriz se utiliza como una matriz de adyacencia donde toma valor de 1 cuando las regiones están contiguas y 0 cuando no lo están.

Ecuación 1: Cálculo del Índice de Moran

$$I = \frac{1}{s^2} * \sum_i \sum_j \frac{(y_i - \bar{y})(y_j - \bar{y})}{\sum_i \sum_j w_{i,j}}$$

- ***Gráficos y mapas de calor***

El cálculo de estas estadísticas descriptivas se basa en generar gráficos de líneas para observar la evolución de las variables que podrían ser utilizadas como complementos al modelo de predicción y en la visualización de mapas de

calor de estas variables en la ciudad de Bucaramanga. También se mencionan los problemas que se encontraron al analizar las variables y cómo podrían afectar los resultados de los modelos de predicción.

3.2. Predicciones

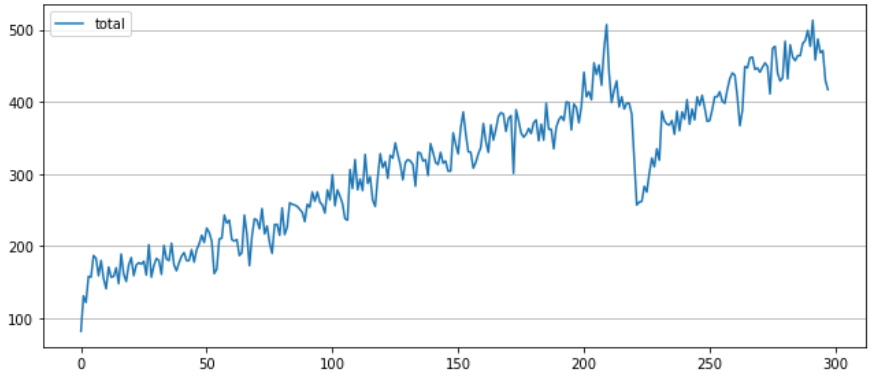
Las predicciones se hicieron de dos maneras generarles. La primera es con la construcción de una matriz de delitos y la segunda con modelos autorregresivos de series de tiempo.

3.2.1. Matriz de delitos

La matriz de delitos se construye a partir de la base de datos de delitos enviada por la Alcaldía de Bucaramanga (homicidios, lesiones y hurtos) y con las infracciones registradas en el Registro Nacional de Medidas Correctivas (RNMC). Todos estos tipos de delitos se calcularon para cada Sección DANE de Bucaramanga de manera semanal, de lunes a domingo. También se incluye en la matriz información de delitos de cada vecino de las Secciones DANE, con el fin de verificar si los delitos pasados de los vecinos influyen en los actuales. Las fechas de los datos, luego del procesamiento y los filtros, comienzan el 2 de enero de 2017 (lunes) hasta el 31 de marzo de 2021 si se incluyen los datos del RNMC, y hasta el 5 de septiembre si no se incluyen.

La Gráfica 1 muestra la evolución de los delitos totales desde la primera semana de 2017 hasta septiembre de 2021. Hay un poco menos de 300 semanas y se ve claramente el choque generado por el COVID-19. Por ello, se hicieron predicciones tomando en cuenta toda la serie y períodos antes y después del COVID-19. Los resultados con las semanas anteriores al COVID-19 fueron los mejores.

Gráfica 1. Delitos totales en Bucaramanga por semana: enero 2017 hasta



Fuente: elaboración propia

Una versión simplificada de la matriz se muestra en la Tabla 1. Allí se muestra un ejemplo hipotético con dos secciones DANE. La columna “Semana” representa las semanas de lunes a domingo que se tienen en cuenta para las variables de delitos. “Secciones” simplemente menciona la Sección DANE que representa la fila. Las columnas de delitos contienen los delitos en de cada semana de la tabla y en momentos diferentes del tiempo. En la tabla original se encuentran por separado los delitos totales, los hurtos, las lesiones personales, los homicidios y las infracciones. La variable “Delitos t” contiene los delitos de la semana t, mientras que “Delitos t-1” y “Delitos t-2” contienen los delitos de las semanas t-1 y t-2, respectivamente. Por último, las columnas “Delito en Sección 1 en t-1” y “Delito en Sección 2 en t-1” muestran si hubo un delito en la Sección vecina en la semana anterior. En el modelo final se utilizan 10 rezagos de cada variable para predecir los delitos en Bucaramanga.

Tabla 1. Matriz de delitos

Semana	Secciones	Delitos t	Delitos t-1	Delitos t-2	Delito en Seccion 1 en t-1	Delito en Seccion 2 en t-1	(…)
1	sección1	1	.	.	.	.	(…)
2	sección1	5	1	.	0	0	(…)
3	sección1	0	5	1	0	1	(…)
4	sección1	2	0	5	0	0	(…)
5	sección1	1	2	0	0	1	(…)
1	sección2	0	.	.	.	.	(…)
2	sección2	0	0	.	1	0	(…)
3	sección2	3	0	0	0	0	(…)
4	sección2	0	3	0	1	0	(…)
5	sección2	2	0	3	1	0	(…)
(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)

Fuente: elaboración propia

A partir de la matriz de delitos se predicen los delitos de cada semana en una Sección DANE de acuerdo con la siguiente función:

Ecuación 2. Modelo general de predicción

$$\begin{aligned} delito_{s,t} = f(&delito_{s,t-1}, delito_{s,t-2}, (\dots), delito_{s,t-n}, \\ &hurtos_{s,t-1}, hurtos_{s,t-2}, (\dots), hurtos_{s,t-n}, \\ &lesiones_{s,t-1}, lesiones_{s,t-2}, lesiones_{s,t-n}, \\ &homicidios_{s,t-1}, homicidios_{s,t-2}, (\dots), homicidios_{s,t-n}, \\ &delito_{v1,t-1}, delito_{v2,t-2}, (\dots), delito_{v2,t-n}) \end{aligned}$$

Donde “s” es la Sección DANE, “t” es el tiempo (semana) y “v” es la Sección DANE de un vecino de “s”. La variable “delito” representa la suma de hurtos, lesiones y homicidios, mientras que “hurtos”, “lesiones”, “homicidios” e “infracciones” son los delitos por separado.

- *Modelos de predicción*

Entre los modelos de predicción se utilizaron XgBoost, *Quadratic Discrimination Analysis*,*Support Vector Machine* y *Random Forest.* El modelo con mejores resultados fue el de XgBoost. Estos modelos se pueden utilizar en el caso de los problemas de clasificación, donde se busca predecir los valores de variables categóricas como las que indican si hay un delito o no en una Sección DANE. En este caso se predice si en una Sección DANE para una semana específica hubo o no un delito de acuerdo con variables independientes que incluyen delitos totales, hurtos, lesiones, homicidios e infracciones con rezagos. También se incluye la información de los delitos rezagados en las Secciones DANE vecinas.

Como los modelos de predicción son supervisados, contienen variables independientes y dependientes, se dividió la base de datos en secciones de entrenamiento y prueba. Estas divisiones se hicieron para períodos pre y post COVID-19, es decir, seleccionando semana anteriores a 2020, por un lado, y semanas posteriores a esta fecha, por el otro.

- *Interpretación de resultados*

Los modelos se interpretan con las métricas de *recall y precision, basados en matrices de confusión*. Como este es un ejercicio de predicción binomial (se predice si va a haber o no un delito), la matriz de confusión tiene el tamaño 2X2 y muestra los resultados de la siguiente manera, suponiendo que los datos que se quieren predecir tienen los valores "positivo" y "negativo" (Tabla 2):

Tabla 2. Matriz de confusión general

		Categoría predicha	
		Negativo	**Positivo**
Categoría Original	**Negativo**	Verdaderos negativos	Falsos positivos
	Positivo	Falsos negativos	Verdaderos positivos

Fuente: elaboración propia

La Tabla 3 muestra la matriz de confusión en el caso de la predicción de proyectos de inversión con potenciales problemas.

Tabla 3. Matriz de confusión para el caso de predicción de delitos

		Categoría predicha	
		No problema	**Problema**
Categoría Original	**No problema**	Predicción correcta: no hay delito	Predicción incorrecta: hay delito
	Problema	Predicción incorrecta: no hay delito	Predicción correcta: hay delito

Fuente: elaboración propia

Las dos métricas que se utilizaron para validar los resultados de las predicciones son las de *recall* y *precision.* Estas se calculan de la siguiente manera, a partir de la información en la matriz de confusión (Tabla 2).

Tabla 4. Métricas de rendimiento de modelos

Métrica	Cálculo
Recall	$\frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos}$
Precision	$\frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos}$

Fuente: elaboración propia

La métrica *recall* calcula el porcentaje de delitos (verdaderos positivos) que sí se lograron predecir entre todos los delitos. Por ejemplo, si se predicen 300 delitos y el total de delitos son 1.000, entonces el *recall* sería 30%. Por último, *precision* mide el porcentaje de delitos que en realidad fueron delitos. Es decir, si un modelo predice 1.000 delitos, pero 500 no fueron, es decir, se predijeron incorrectamente como delitos (son falsos positivos), entonces el *precision* sería 50%.

Saber cuál es el mejor modelo estimado depende entonces del *recall* y *precision*. El mejor modelo sería uno con *recall* de 100% y *precision* de 100%. Sin embargo, estas dos métricas casi siempre entran en conflicto. Si se logra predecir

una gran cantidad de delitos (se tendría un *recall* alto), es muy probable que se tenga una cantidad alta de falsos positivos, por lo cual disminuiría el *precision*. Por esta razón, la decisión final del modelo utilizado para la predicción depende si se tolera tener una cantidad alta de predicción delitos con muchos falsos positivos o si se desea tener mejor una cantidad baja de delitos predichos con más exactitud.

- *Umbral de éxito*

Un aspecto importante de señalar es el del umbral de éxito de los modelos de clasificación. El siguiente ejemplo se usará para explicar cómo funciona el umbral. La Tabla 5 muestra un resultado hipotético con los resultados de la predicción de delitos.

Tabla 5. Resultado hipotético de predicción de delitos

Sección DANE	Probabilidad de delito
1	0.674
2	0.019
3	0.283
4	0.005
5	0.562

Fuente: elaboración propia

La columna "Probabilidad de delito" contiene las probabilidades que haya un delito en una Sección DANE. Como se trata de un modelo de clasificación, se tiene que definir cuándo se considera que haya un delito o no en una Sección DANE. Por ello se supone una probabilidad mínima de éxito, o de que haya un delito. Usualmente la probabilidad mínima es 0.5 y en este caso, con esta probabilidad, se predeciría que las Secciones 1 y 5 tendrían delitos. Sin embargo, este umbral puede aumentar para considerar únicamente como delitos los casos con las probabilidades más altas. Por ejemplo, si el umbral mínimo es de 0.6, entonces solo la Sección 1 sería predicha con un delito.

Es importante mencionar el umbral de éxito para la predicción porque al aumentarlo se tienen en cuenta las Secciones DANE con probabilidades más altas de tener delitos, lo cual incrementa la métrica *precision* pero disminuye el *recall*. Es decir, al aumentar el umbral de éxito se predicen menos delitos correctamente, pero los delitos predichos tienen menos falsos positivos en sus resultados. Los resultados se presentan con distintos valores del umbral de éxito para verificar la variación en las métricas *recall* y *precision*.

3.2..2 Modelos autorregresivos

El segundo modelo utilizado consiste en predecir los delitos de varias semanas a partir de los rezagos de los delitos y de rezagos de variables explicativas. Para ello se utilizó la librería Skforecast, la cual permite hacer predicciones de series de tiempo incluyendo variables independientes y modelos de machines learning de Scikit-Learn, una de las librerías de Python para predicciones de aprendizaje de máquina. A partir de las variables se predicen recursivamente los valores subsiguientes, es decir, que cada nueva predicción se utiliza como insumo para la predicción siguiente. La fuente literaria para estos modelos se encuentra en el siguiente enlace web:

https://www.cienciadedatos.net/documentos/py27-forecasting-series-temporales-python-scikitlearn.html

Por lo tanto, se entrena un modelo para cada Sección DANE, que incluye las variables de delitos e infracciones de esa sección y también el número de delitos de los vecinos. Se calcularon modelos con 50 y 100 rezagos. Adicionalmente,

con el fin de mejorar los resultados, se hicieron predicciones para los delitos de Sectores DANE, es decir, con una mayor agregación que contenga más información de delitos.

Para medir los resultados de estos modelos se utiliza la métrica del error cuadrático medio (*MSE,* en inglés) y se observan las gráficas de las predicciones sobre la serie de tiempo. El MSE se calcula de acuerdo con la siguiente ecuación:

Ecuación 3. Error cuadrático medio

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Done N es el número total de observaciones, Y_i es el valor i de la serie original y $\hat{Y}_i$ es el valor i predicho. Por lo tanto, el MSE es el promedio de las diferencias entre el valor original de la serie y el valor predicho elevadas al cuadrado.

3.3. Herramienta de visualización

La metodología consiste en utilizar las bases de datos de delitos históricos y predichos junto con shapefiles geográficos desagregados por Sección DANE de Bucaramanga y la librería Streamlit de Python para crear una herramienta de visualización con un mapa de calor. Esta permite seleccionar una fecha y mostrar los delitos históricos o predichos de la semana, de lunes a domingo, de cada Sección DANE de la ciudad.

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación. Toda retroalimentación desde un punto de vista experto o de usuario por parte de la DSJG es bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

4.1. Estadísticas descriptivas

Los resultados se presentan para cada una de las bases de datos enviada por la Alcaldía de Bucaramanga que pudo ser analizada (hay varios recursos enviados que no tienen un formato codificable) y para la base de delitos de la Policía Nacional para los años 2016-2019.

4.1.1. Policía Nacional – delitos en Bucaramanga 2014-2019

En general, se evidencia que los delitos más comunes son los hurtos (55,5%), seguido por las lesiones personales (23,6%), la violencia intrafamiliar (19,7) y, por último, los homicidios (1,1%). Adicionalmente, los homicidios y lesiones personales tienen una tendencia más o menos estable. Sin embargo, la violencia intrafamiliar tiende a disminuir entre 2014 y 2019, pero los hurtos tienen un incremento bastante marcado en este mismo período. Por último, todos los tipos de delitos en Bucaramanga tienen una alta concentración en ciertos lugares geográficos, especialmente en el centro de la ciudad. Para más información sobre esta base de datos, por favor leer el documento "Entregable_descriptivos_2020.pdf" del proyecto de predicción de delitos en Bucaramanga de 2020 que se envió junto al Entregable 1 del proyecto, que es el mismo del año 2020.

4.1.2. Índice de Moran

La Tabla 6 muestra los resultados del Índice de Moran y su nivel de significancia sobre la base de datos de delitos 2016-2019. En ella se ve que sí existe autocorrelación espacial únicamente con el nivel de delitos agregados (considerando cualquier tipo de delito sin discriminar qué tipo). Por su parte, al discriminar por delito, no se evidencia autocorrelación espacial. Esto se puede deber a que existe una relación espacial de un delito frente a otro delito y no precisamente dentro del mismo tipo de delito (para el delito de homicidios no fue posible calcular el índice de Moran dado que no existe mucha variabilidad en esta variable).

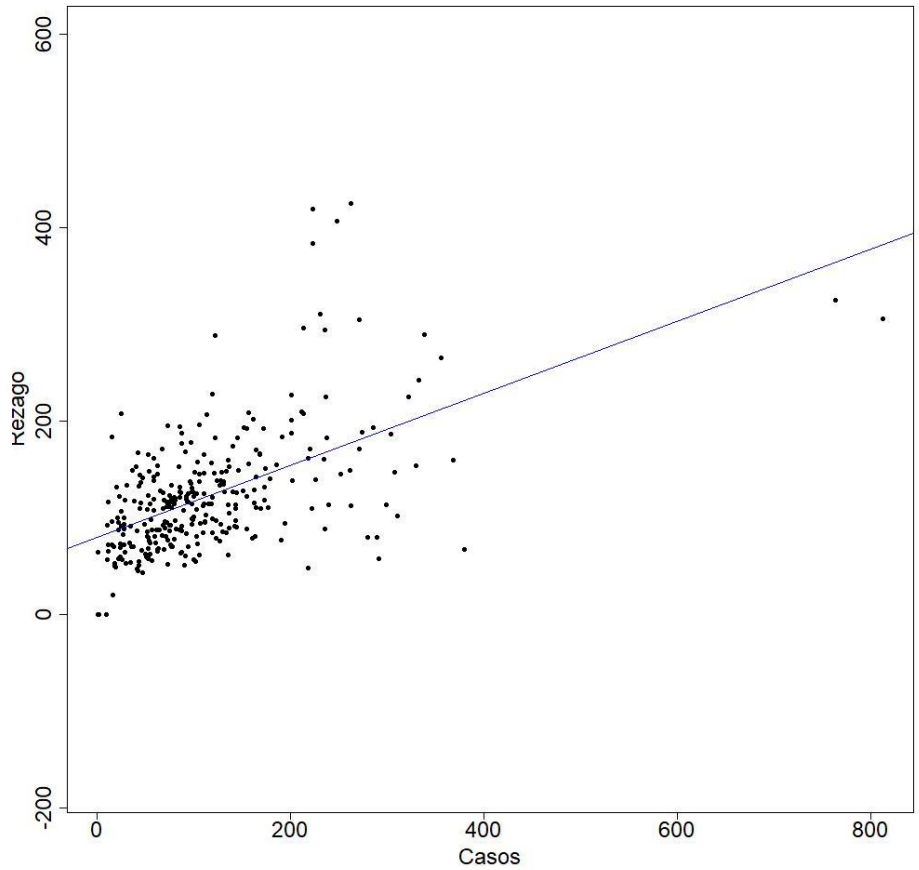
Tabla 6: Índice de Moran para cada delito y agregado en Bucaramanga

Delito	Índice de Morán	P-value
Todos	0,3543	0,0010
Homicidio	No disponible	No disponible
Hurto a personas	-0,0033	0,5190
Lesiones personales	-0,0036	0,6635
Violencia intrafamiliar	-0,0016	0,0960

Fuente: Cálculos propios

Por su lado, Gráfica 2 muestra los delitos agregados en función a sus rezagos espaciales, es decir, es la relación entre la cantidad de delitos para cada sección contra la cantidad de delitos en sus secciones contiguas (vecinas). Lo anterior implica que hay tantos rezagos como polígonos de distancia pueda haber frente a cada polígono. Si el gráfico de dispersión muestra una tendencia positiva, quiere decir que existe una correlación espacial positiva, si es negativa, implica una correlación espacial negativa y si es cero, significa que no hay correlación espacial. En el caso de los delitos agregados sí existe una relación positiva espacial con sus rezagos, de tal modo que un aumento en el nivel general de delitos en una sección llevaría a un aumento general de delitos en regiones contiguas.

Gráfica 2: Delitos agregados en función a sus rezagos

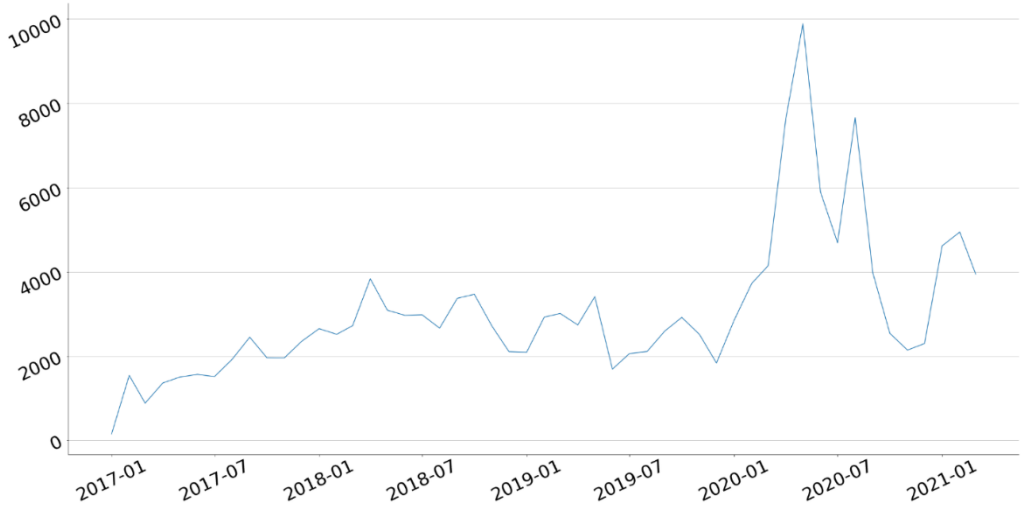


Fuente: Cálculos propios

4.1.3. Registro Nacional de Medidas Correctivas (RNMC)

La información del RNMC se encuentra desde el 1 de enero de 2017 hasta el 31 de marzo de 2021. La Gráfica 3 muestra la evolución del total de las medidas aplicadas en la ciudad de Bucaramanga durante este periodo. Si bien la tendencia general es positiva, con un incremento desde enero de 2017 y estabilización hasta finales de 2019 alrededor de los 3.000 casos mensuales, a partir de 2020 hubo una explosión en los casos reportados y luego un abrupto descenso en la segunda mitad del año.

Gráfica 3. Medidas Correctivas – enero 2017 hasta marzo 2021 (serie mensual)

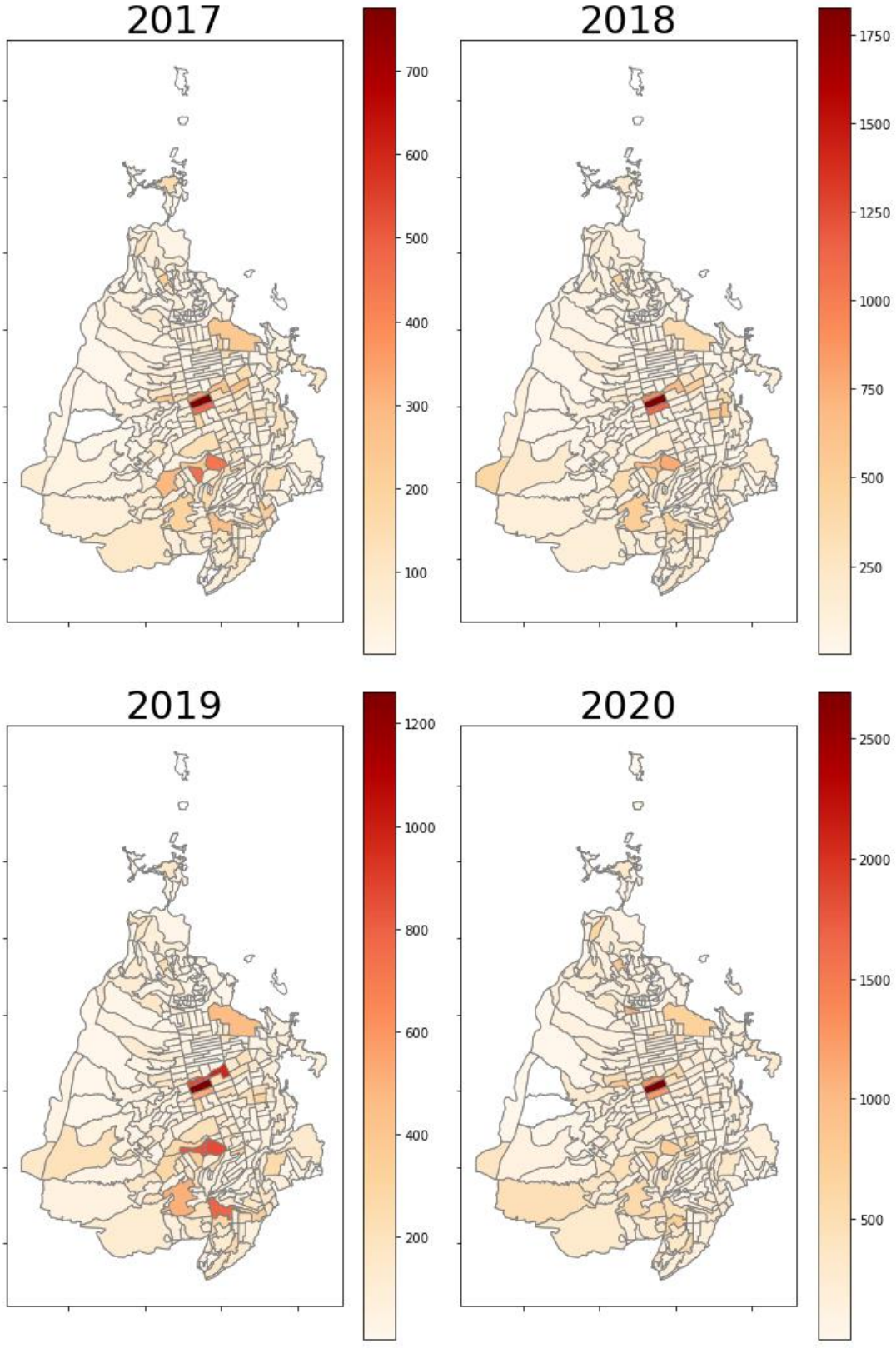


Fuente: RNMC

Los mapas de calor de las medidas correctivas aplicadas en desde 2017 hasta 2020 muestran que tienen una alta concentración en algunas Secciones DANE del centro de la ciudad. En todo caso, hay casos registrados en la gran mayoría de Secciones DANE. La 0 muestra los mapas de calor de las medidas correctivas aplicadas en Bucaramanga en los años 2017 hasta 2020.

En la Gráfica 13, en los anexos, se presentan los mismos mapas de calor que se acaban de mencionar con la diferencia que se dividen las medidas aplicadas por el kilómetro cuadrado de cada Sección DANE.

Gráfica 4 Medidas correctivas – mapas de calor

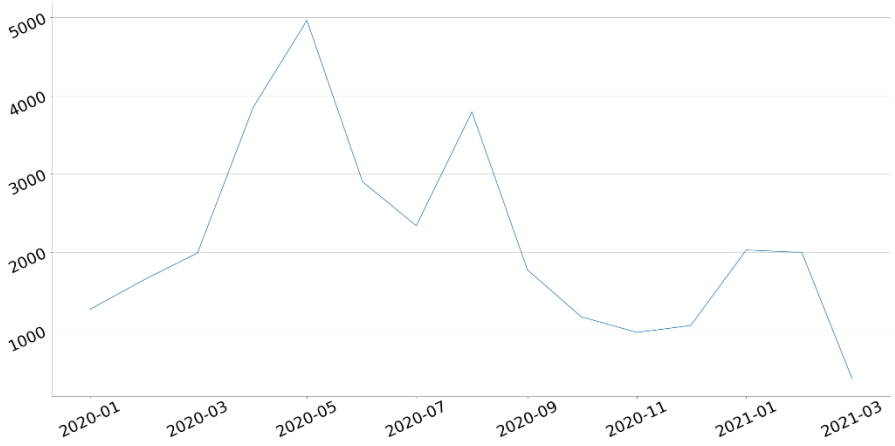


Fuente: RNMC

4.1.4. Código Nacional de Seguridad y Convivencia Ciudadana (CNSCC)

Los datos del CNSCC se encuentran disponibles desde enero de 2020 hasta marzo de 2021 y tienen una tendencia negativa durante este período a partir de mayo de 2020. En este mes se encontraron en su punto máximo de la serie, cerca de 5.000 casos reportados y disminuyeron hasta encontrarse en valores cercanos a los 2.000 casos en febrero de 2021 (marzo de 2021 cuenta únicamente con datos hasta el día 8). La Gráfica 5. muestra los casos del CNSCC mensuales.

Gráfica 5. Infracciones del CNSCC – enero 2020 hasta marzo 2021 (serie mensual)

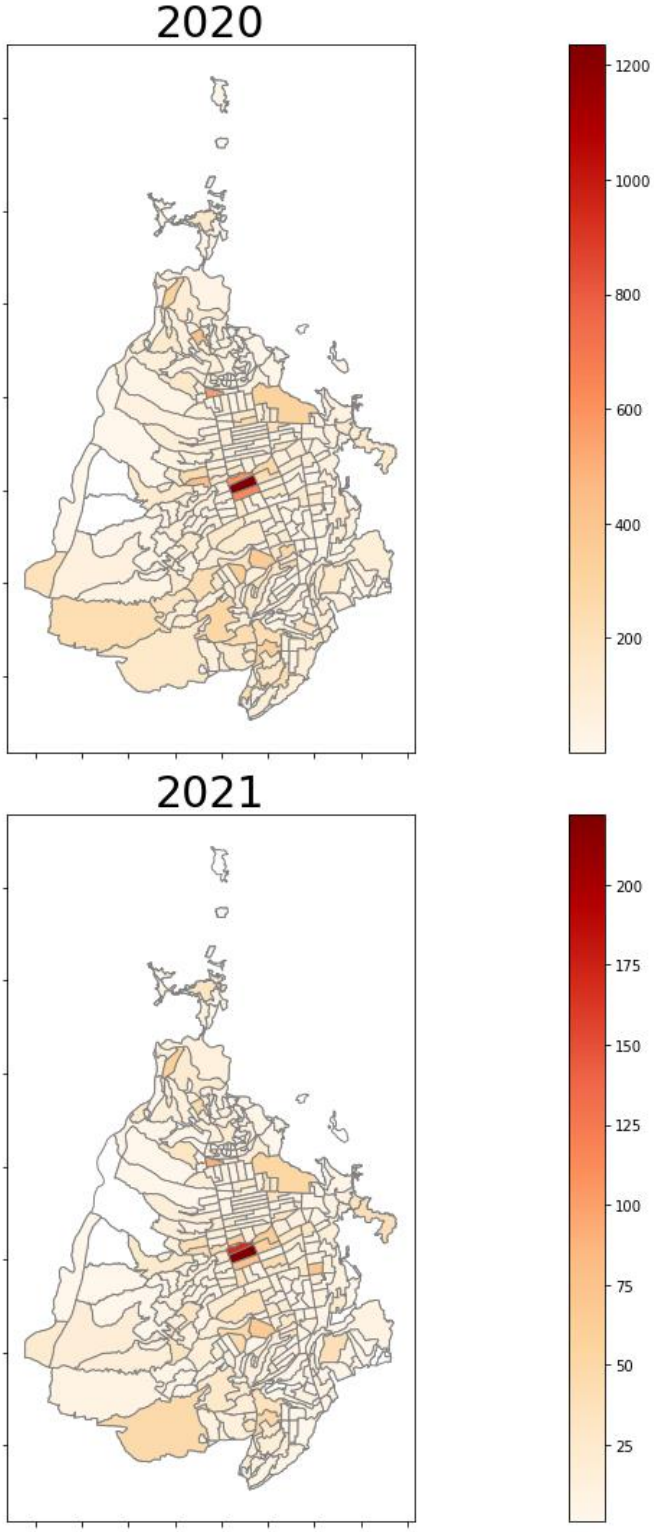


Fuente: CNSCC

La Gráfica 6 muestra un mapa de calor de los casos reportados del CNSCC en Bucaramanga para cada año de la muestra. La característica más notable, al igual que las infracciones del RNMC, es la alta concentración de casos unas cuantas Secciones DANE en el centro de la ciudad, si bien la gran mayoría de Secciones tuvieron intervenciones de la Policía.

En la Gráfica 14, en los anexos, se encuentran los mismos mapas de calor, pero con los casos reportados de cada Sección DANE divididos por los kilómetros cuadrados de cada Sección. De esta manera se observa que el centro de la ciudad sigue siendo el lugar con más casos reportados del CNSCC, sin embargo aparecen otros puntos importantes, especialmente al norte de la ciudad.

Gráfica 6: Distribución de infracciones del CNSCC en Bucaramanga para el 2020 y 2021

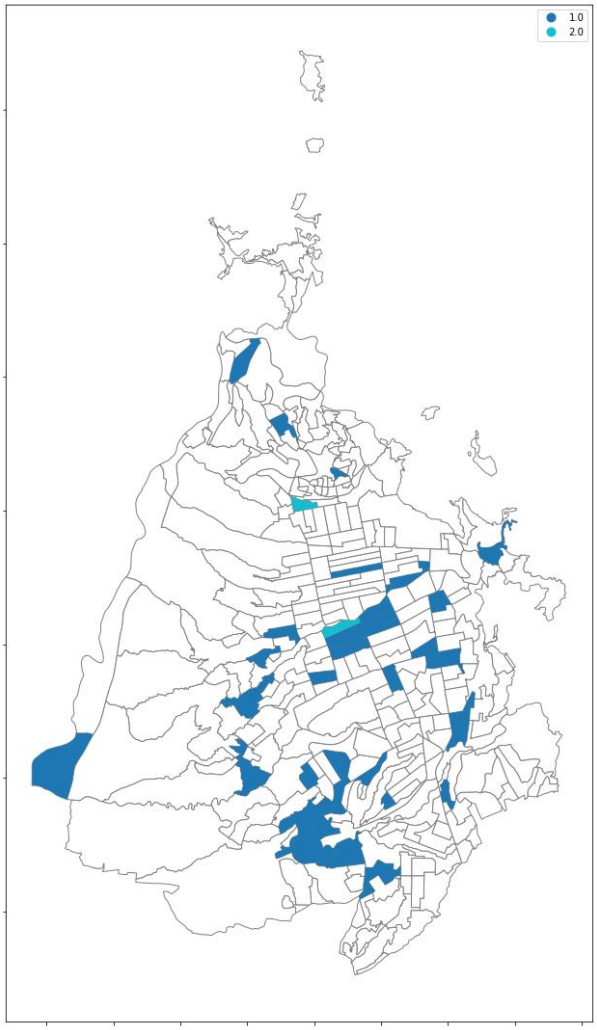


Fuente: CNSCC

4.1.5. Unidades de policía

La información enviada por la Alcaldía de Bucaramanga cuenta con la ubicación geográfica de 315 unidades de policía, entre los cuales se encuentran CAI, estaciones, fuertes y comandos. No se especifica la fecha de operación de cada unidad. Las secciones DANE que cuentan con unidades de policía se observan en la Gráfica 7.

Gráfica 7 Unidades de policía – Secciones DANE

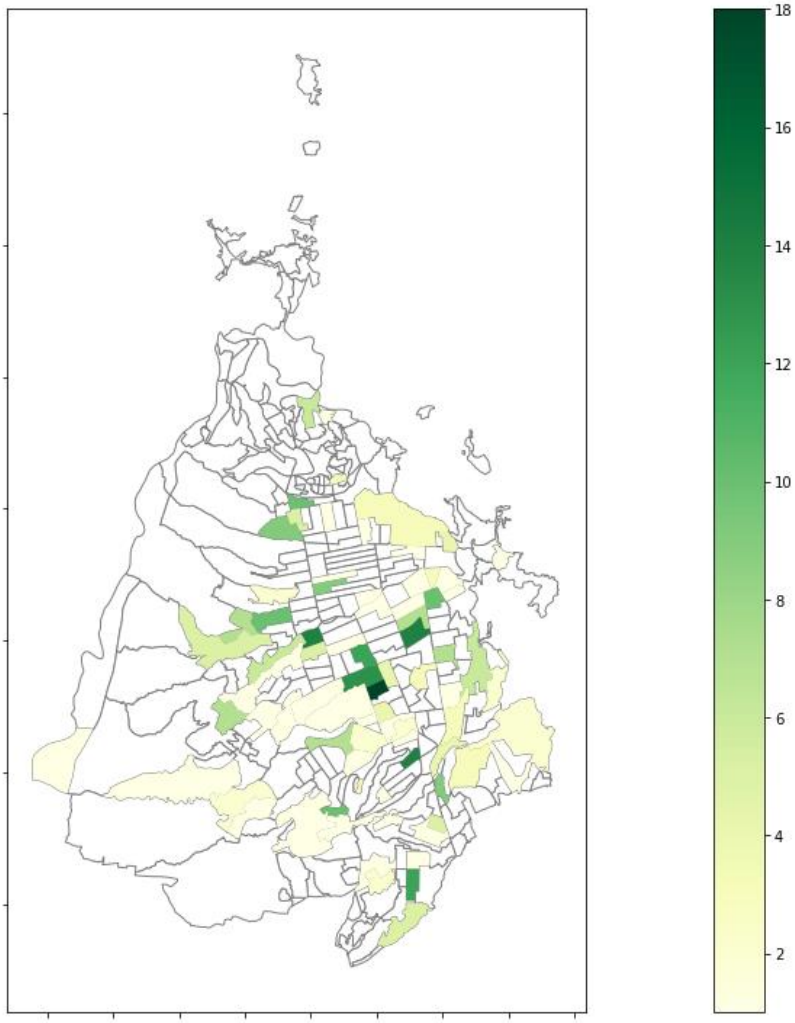


Fuente: unidades policiales MEBUC

4.1.6. Cámaras de seguridad

La tabla de cámaras de seguridad no cuenta con una localización geográfica exacta, sino con las direcciones de las cámaras en formato texto. Por eso es necesario obtener las coordenadas geográficas mediante programas especializados. Sin embargo, varias direcciones no son lo suficientemente exactas y no es posible obtener la totalidad de las ubicaciones en el formato latitud-longitud. Por esa razón, del total de 416 cámaras de seguridad reportadas en el conjunto de datos, se pudo extraer la ubicación exacta de 331 cámaras. Adicionalmente, no se especifica la fecha de instalación de cada cámara. La Gráfica 8 muestra el mapa de calor de las 331 cámaras de seguridad en Bucaramanga.

Gráfica 8 Cámaras de seguridad



Fuente: Cámaras de seguridad – Alcaldía de Bucaramanga

4.2. Modelos de predicción

4.2.1. Matriz de delitos

A continuación se muestran los resultados para el modelo XgBoost de la metodología presentada en la Sección 0. La Tabla 7 contiene los resultados para el modelo entrenado con semanas anteriores a 2020, es decir, sin incluir los efectos ocasionados por el COVID-19. La Tabla 8 contiene los resultados con el modelo entrenado con semanas de 2020 y 2021 (hasta marzo).

Los resultados de la Tabla 7 muestran que el *recall* máximo es de 0.411 (se predicen 41.1% de todos los delitos correctamente) si el umbral de éxito es de 0.5 y disminuye significativamente si sube el umbral. El *precision* es de 0.638 (63.8% de los delitos predichos son verdaderos positivos) con el umbral de 0.5. Si bien el *precision* aumenta significativamente con la subida del umbral, el *recall* disminuye demasiado hasta el punto de que con el umbral de 0.9 solo se predicen un poco más de 5% de los delitos, si bien con una precisión alta.

Tabla 7. Resultados pre COVID-19

Umbral de éxito	Recall	Precision
0.5	0,411	0,638
0.6	0,288	0,705
0.7	0,186	0,786
0.8	0,099	0,874
0.9	0,052	0,961

Fuente: elaboración propia

La Tabla 8 muestra el mismo modelo si se entrena con las semanas de la segunda mitad de 2020 hasta marzo de 2021. Claramente disminuye el rendimiento, dado que para el umbral de éxito de 0.5 el *recall* es es 0.26 y el *precision* de 0.59.

Tabla 8. Resultados post COVID-19

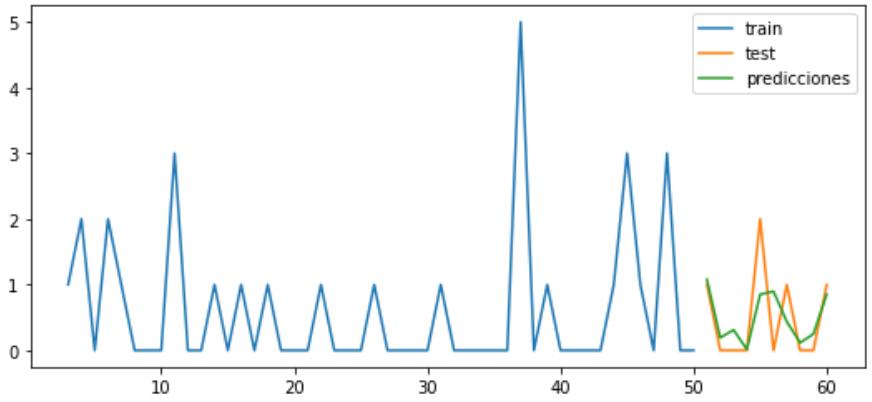
Umbral de éxito	Recall	Precision
0.5	0,324	0,596
0.6	0,197	0,689
0.7	0,102	0,764
0.8	0,045	0,882
0.9	0,005	1,000

Fuente: elaboración propia

4.2.2. Series de tiempo

En general, se encontró que cada Sección DANE contiene muy pocas observaciones para que las predicciones por series de tiempo sean efectivas. La Gráfica 9 muestra una predicción en una Sección donde se encuentran muchas observaciones que son 0.

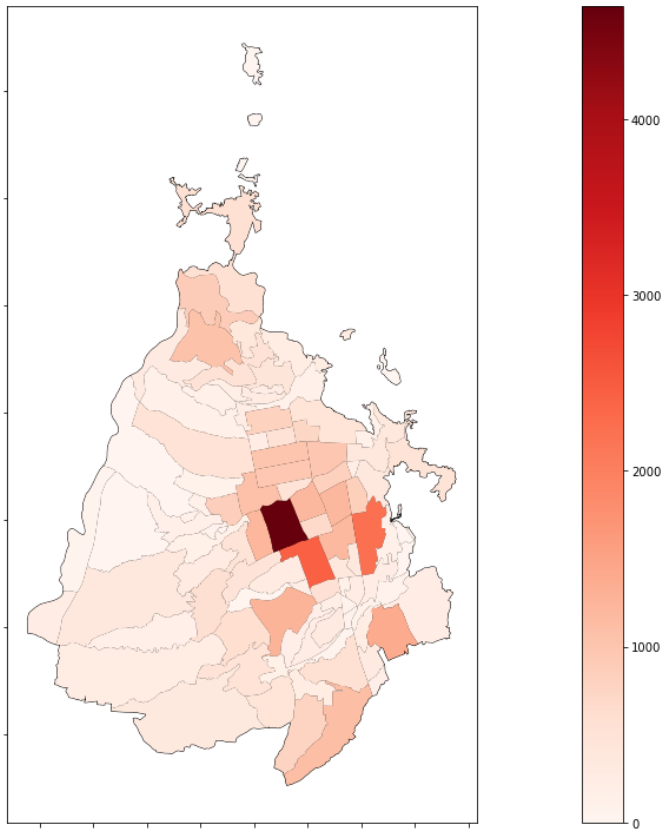
Gráfica 9.Delitos predichos para una Sección DANE (pre COVID-19) por semana



Fuente: elaboración propia

Por esta razón se intentó hacer las predicciones por Sector DANE, que tienen un nivel de agregación por encima de las Secciones DANE. La Gráfica 10 muestra los delitos totales ocurridos entre 2016 y 2021 en Bucaramanga por Sector DANE.

Gráfica 10. Delitos por sector DANE



Fuente: elaboración propia

En todo caso, este nivel de desagregación sigue siendo muy bajo para poder aplicar la metodología de series de tiempo. Los resultados de las predicciones se encuentran en la Tabla 1. Si bien parecen bajos, y por lo tanto buenos, como hay tantos sectores con pocos delitos las predicciones tienden a acercarse a cero, lo cual influye en el MSE.

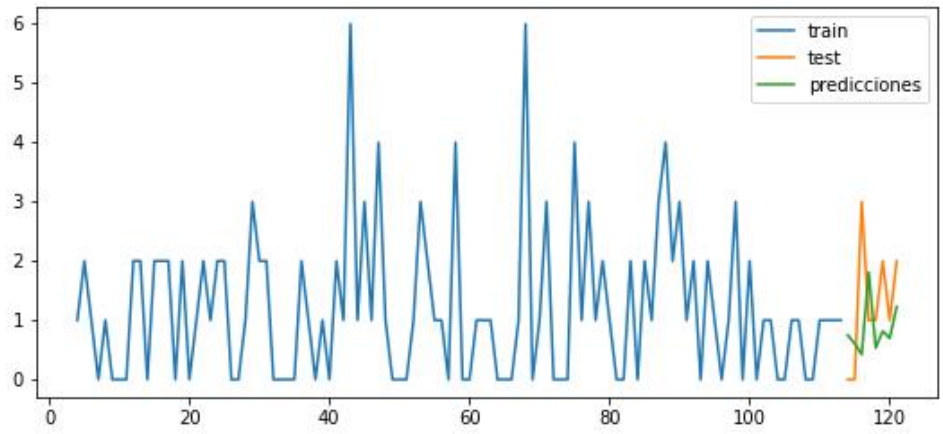
Tabla 9. Métricas MSE

Modelo	Número de zonas geográficas	MSE
RandomForest	Todos los sectores	0,578
RandomForest	20 Sectores con más delitos	1,071
XgBoost	Todos los sectores	0,504
XgBoost	20 Sectores con más delitos	0,857

Fuente: elaboración propia

La Gráfica 11 muestra un ejemplo con un Sector DANE que cuenta con varios delitos. En este caso podría justificarse la predicción con series de tiempo.

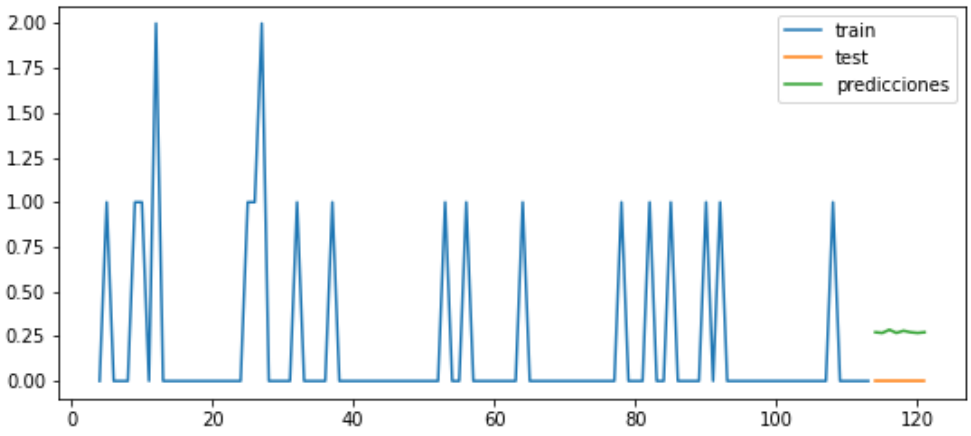
Gráfica 11. Predicción para Sector DANE del total de delitos por semana – ejemplo 1 (pre COVID-19)



Fuente: elaboración propia

Sin embargo, la mayoría de Sectores DANE no cuentan con muchos delitos, por lo que la generalización de los resultados no fue satisfactoria. La Gráfica 12 muestra un Sector DANE sin muchos delitos y su predicción.

Gráfica 12. Predicción para Sector DANE del total de delitos por semana – ejemplo 2 (pre COVID-19)



Fuente: elaboración propia

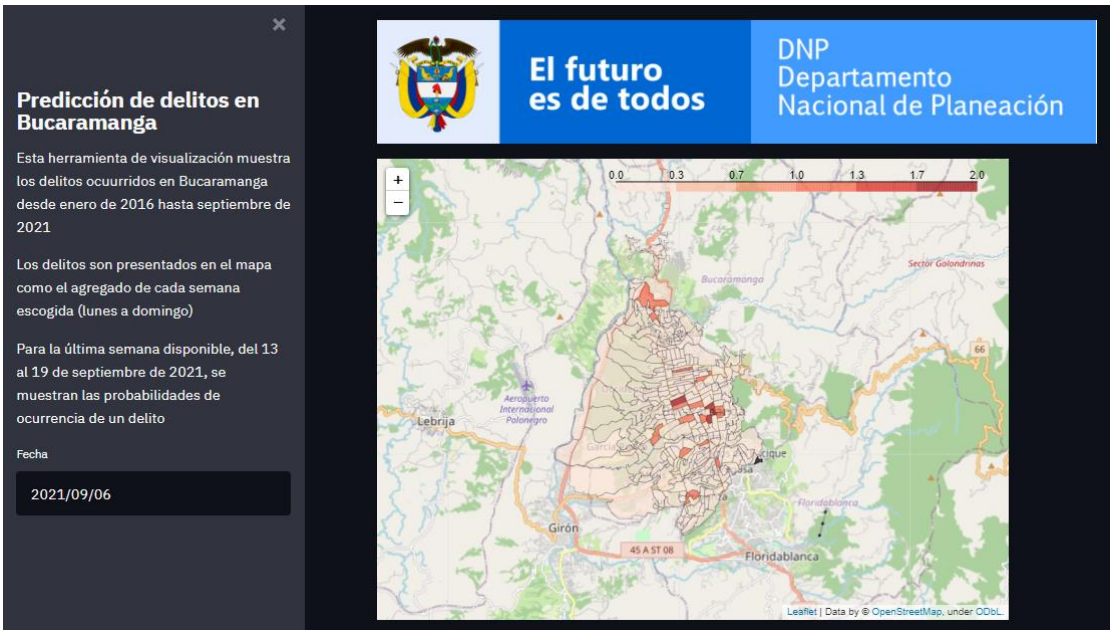
4.3. Herramienta de visualización

Le herramienta de visualización se desarrolló en Python con la librería Streamlit y los datos de delitos históricos y predichos como insumo para ser visualizados por semana y Sección DANE en la ciudad de Bucaramanga.

4.3.1. Primera vista de la herramienta de visualización

Al ejecutar el código para abrir la herramienta, se abrirá un navegador de internet y se verá lo siguiente.

Ilustración 1. Primera vista de la herramienta de visualización



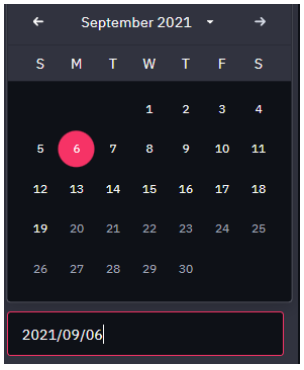
Fuente: herramienta de visualización

La barra lateral izquierda contiene una descripción breve de la herramienta y la opción de escogencia de una fecha desde enero de 2016 hasta septiembre de 2021. El mapa de Bucaramanga se encuentra a la derecha de esta barra y ahí se pueden ver la información de delitos de la ciudad para la semana (de lunes a domingo) escogida.

4.3.2. Selección de fecha

Debajo del texto "Fecha" de la barra lateral izquierda se encuentra el calendario de fechas, del cual se puede seleccionar un día entre el 4 de enero de 2016 y el 19 de septiembre de 2021 para mostrar la información de delitos en Bucaramanga de la semana del día escogido (de lunes a domingo). El calendario se ve de la siguiente manera, una vez se presiona sobre la opción.

Ilustración 2. Selección de fechas

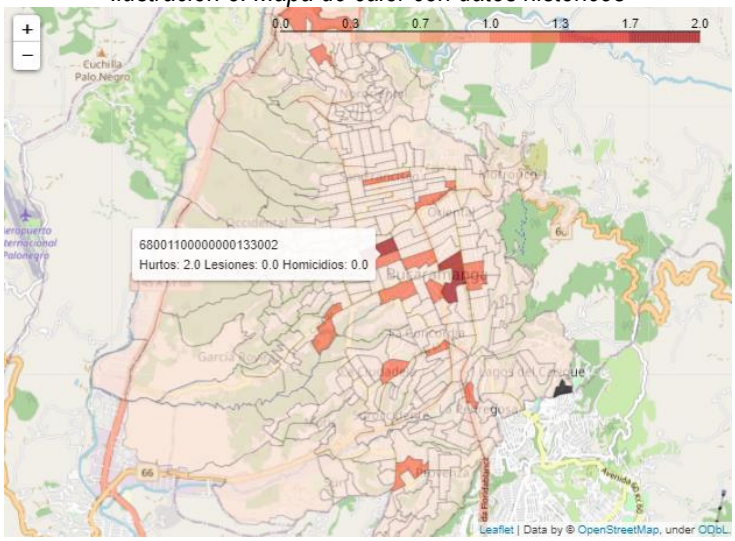


Fuente: herramienta de visualización

4.3.3. Vista del mapa con una fecha específica de datos históricos

El mapa interactivo muestra información de delitos en Bucaramanga para cada Sección DANE de una semana específica, escogido en el calendario de la barra lateral izquierda. En caso de escoger una fecha anterior a la última semana disponible, del 13 al 19 de septiembre de 2021, se visualizarán los datos históricos de hurtos, lesiones y homicidios en cada Sección DANE. La Ilustración 3 muestra un ejemplo de cómo ve el mapa cuando se pasa el cursor sobre una Sección DANE que tuvo 2 hurtos, 0 lesiones y 0 homicidios en la semana escogida.

Ilustración 3. Mapa de calor con datos históricos



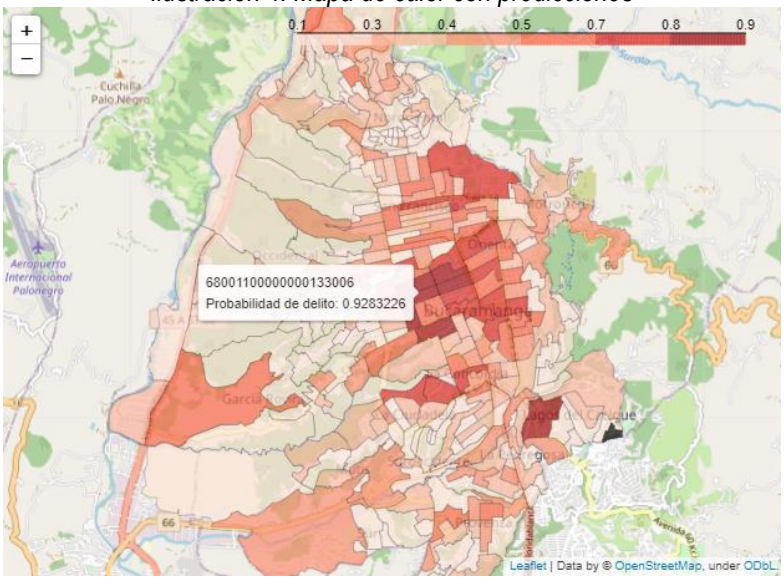
Fuente: herramienta de visualización

Los colores del mapa de calor varían desde blanco hasta rojo oscuro. Entre más oscura se encuentre una Sección DANE, significa que más delitos hubo en esa Sección.

4.3.4. Vista del mapa con las predicciones de delitos

Si se selecciona en el calendario un día entre el 13 y 19 de septiembre de 2021, se mostrarán en el mapa de calor la probabilidad en cada Sección DANE de que haya un delito. La Ilustración 4 muestra este caso.

Ilustración 4. Mapa de calor con predicciones



Fuente: herramienta de visualización

Los colores del mapa de calor varían desde blanco hasta rojo oscuro. Entre más oscura se encuentre una Sección DANE, significa que mayor probabilidad hay de que suceda un delito.

5. Conclusiones y recomendaciones

A partir de la metodología desarrollada y de los resultados obtenidos para cumplir los objetivos de este proyecto, planteados en el plan de trabajo, se presentan a continuación las principales conclusiones obtenidas por el equipo de la UCD y las principales recomendaciones para un mejor uso y aprovechamiento del proyecto.

1. Tanto las variables de delitos como las de infracciones menores tienen una alta concentración en lugares céntricos de la ciudad de Bucaramanga
2. Con el índice de Moran, se evidenció que los delitos totales en la ciudad de Bucaramanga no son independientes dentro de cada Sección DANE y que lo efectos de los vecinos afectan lo sucedido dentro de un lugar geográfico. Sin embargo, no existe evidencia para afirmar que existe autocorrelación espacial individualmente en cada tipo de delito.
3. Para los modelos de predicción se utilizaron las variables de delitos e infracciones con información histórica y geolocalización.
4. Se recomienda entrenar y estudiar más modelos de predicción para emplear una política de prevención del delito en Bucaramanga. Sería necesario intentar con otros modelos para mejorar los resultados o modificar el modelo de la matriz de delitos. No se recomienda continuar con el modelo de series de tiempo para desagregaciones geográficas altas.
5. Las predicciones de delitos empeoran si se incluyen las semanas correspondientes a 2020 y 2021. Esto probablemente sucede a causa de los choques estructurales ocasionados por el COVID-19 en la ciudad.
6. Dado el choque exógeno generado por el COVID-19 en los datos de delitos, se recomienda esperar un momento prudente de tiempo antes de intentar hacer las predicciones para períodos más recientes
7. Podría considerarse generar agregaciones mayores a las Secciones DANE para hacer las predicciones con mejores resultados
8. Las predicciones por series de tiempo funcionan mejor para agregaciones geográficas mayores. Podría incluso utilizarlas para hacer predicciones del total de delitos en Bucaramanga
9. La herramienta de visualización permite estudiar la información histórica de delitos en Bucaramanga por Sección DANE. Más específicamente, muestra los hurtos, lesiones personales y homicidios.
10. Es necesario seguir las instrucciones del manual de uso, entregado a la DSJG, para garantizar el buen funcionamiento de la herramienta de visualización.

6. Socialización

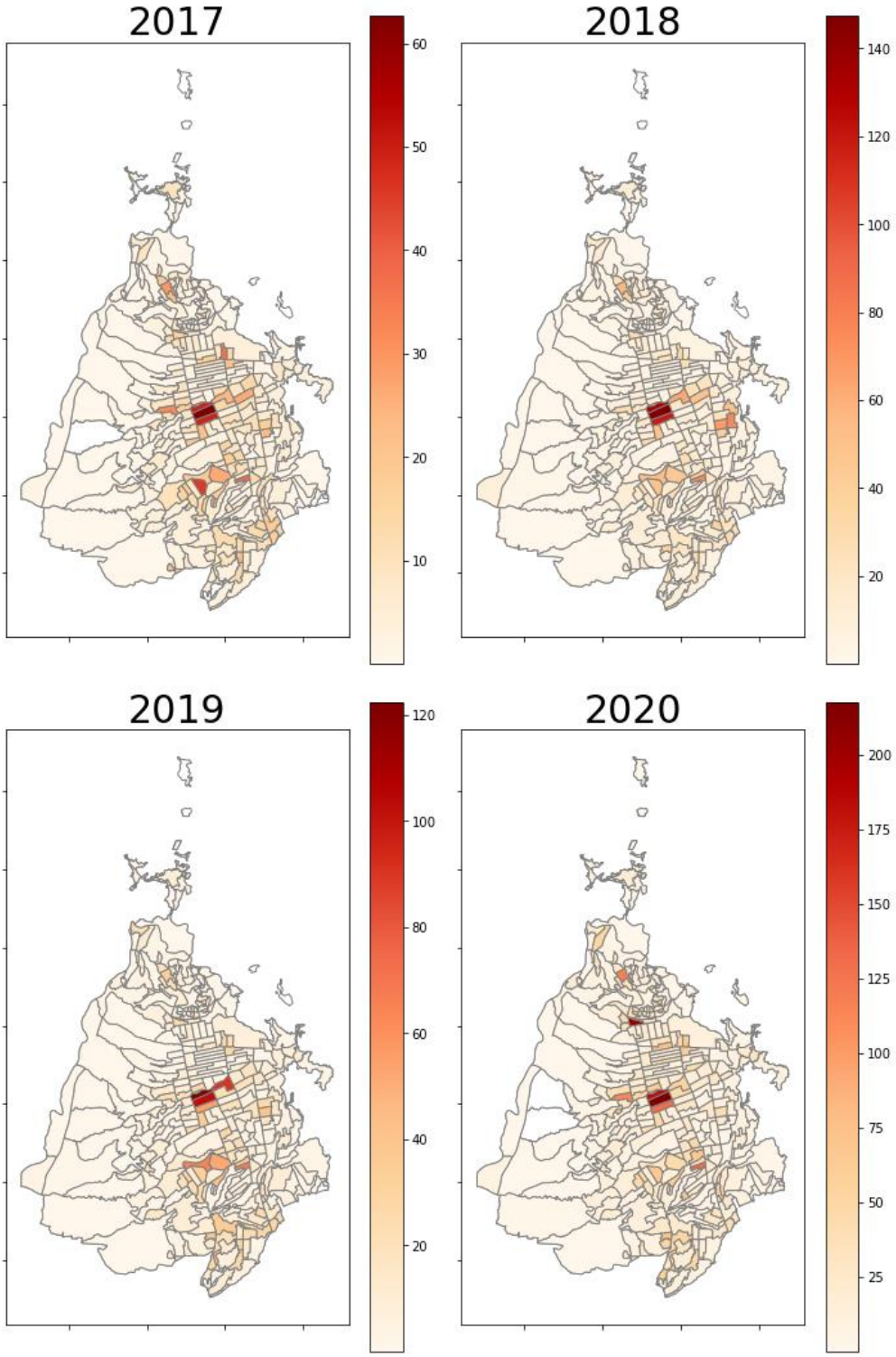
Los resultados fueron compartidos con los miembros de la DSJG y se entregó la herramienta de visualización con el manual de usuario.

Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.

ANEXOS

Gráfica 13. Distribución de medidas correctivas.
Medidas reportadas / kilómetro cuadrado



Gráfica 14. Distribución de infracciones del CNSCC en Bucaramanga para el 2020 y 2021
Infracciones reportadas / kilómetro cuadrado